CAB D0003839.A2 / Final
July 2001

Follow-on Analysis of PAY 97 Test Scores

William H. Sims • Catherine M. Hiatt

4825 Mark Center Drive • Alexandria, Virginia 22311-1850

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

1. AGENCY USE ONLY (*Leave Blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2001	Final

4. TITLE AND SUBTITLE

Follow-on Analysis of PAY 97 Test Scores

5. FUNDING NUMBERS

N00014-00-D-0700

PE - 65154N

PR - R0148

6. AUTHOR(S)

WH Sims, CM Hiatt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center for Naval Analyses
4825 Mark Center Drive
Alexandria, Virginia 22311-1850

8. PERFORMING ORGANIZATION REPORT NUMBER

CAB D0003839.A2

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Manpower Data Center, Monterey CA

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT

Distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

In 1997, in a joint effort, the Department of Labor (DOL) and the Department of Defense (DOD) collected aptitude test data from the Armed Services Vocational Aptitude Battery (ASVAB) on a nationally representative sample of youth. The tests were administered as part of the National Longitudinal Survey of Youth (NLSY97). A subset of data pertaining to youth 18 to 23 years of age is referred to as the Profile of American Youth (PAY 97). In 1999, CNA conducted an initial analysis of PAY 97 tests scores. We concluded that the data sample was missing a large number of persons likely to deplete both the upper and lower levels of aptitude disbributions. We further concluded that the loss would bias resulting norms unless corrected. We recommend weighting the data by race, gender, age, respondent's education, and a proxy for social economic status in an effort to correct the bias. The data were subsequently weighted by NORC. This report describes the follow-on analysis that we conducted on PAY 97 test scores. This work was funded by the Defense Manpower Data Center (DMDC).

14. SUBJECT TERMS

Aptitude tests, aptitudes, AFQT (Armed Forces Qualification Tests), ASVAB (Armed Services Vocational Aptitude Battery), demography, normalizing (statistics), PAY 97 (profile of american youth), recruits, scoring, statistical analysis, statistical distribution, youth

15. NUMBER OF PAGES
72

16. PRICE CODE

17. LIMITATION OF ABSTRACT
SAR

18. SECURITY CLASSIFICATION OF REPORT	19. SECURITY CLASSIFICATION OF THIS PAGE	20. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified	Unclassified	Unclassified

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
299-01

Follow-on Analysis of PAY 97 Test Scores

William H. Sims and Catherine M. Hiatt
Center for Naval Analyses

July 2001

In 1997, in a joint effort, the Department of Labor (DOL) and the Department of Defense (DOD) collected aptitude test data from the Armed Services Vocational Aptitude Battery (ASVAB) on a nationally representative sample of youth. The tests were administered as part of the National Longitudinal Survey of Youth (NLSY97). A subset of data pertaining to youth 18 to 23 years of age is referred to as the Profile of American Youth (PAY 97).

In 1999 the Center for Naval Analyses (CNA) conducted an initial analysis of PAY 97 test scores.[1] We concluded that the data sample was missing a large number of persons likely to deplete both the upper and lower levels of aptitude distributions. We further concluded that the loss would bias resulting norms unless corrected. We recommended weighting the data by race, gender, age, respondent's education, and a proxy for social economic status in an effort to correct the bias.

The data were subsequently weighted by NORC.

This report describes the follow-on analysis that we conducted on PAY 97 test scores. This work was funded by the Defense Manpower Data Center (DMDC).

1. William H. Sims and Catherine M. Hiatt. *Analysis of NLSY97 Test Scores*, Jul 1999, Center for Naval Analyses (CAB 99-66).

Summary

- Current weights for the PAY 97 data are not satisfactory
  - They should be fixed
- We estimate that the mean AFQT for PAY 97 is 51.4. This is not statistically different from the mean found in PAY 80
  - If this estimate is confirmed, it may not be necessary to change the 1980 score scale

The major conclusions from this analysis are:

- Current weights for the PAY 97 data are not satisfactory. Although it may be difficult because of small sample sizes, an attempt should be made to fix the weights.
- We estimated that the mean AFQT for PAY 97 is 51.4. This is not statistically different from the mean found in PAY 80. If this estimate is confirmed, it may not be necessary to change the 1980 score scale. Changing norms and changing score scales are two different actions. We considered current norms for a population group to be the score distribution of that group on *any* score scale. We viewed the construction of a new score scale from that score distribution as a separate action.

Topics

- AFQT and month tested
- Estimates of AFQT from external benchmarks
- Estimates of AFQT from PAY 97
- Are the current weights satisfactory?
- Design effect
- Equivalence of PAY 80 and PAY 97 AFQT means
- AFQT by subgroup

The topics covered in this report are:

- AFQT and month tested
- Estimates of AFQT from external benchmarks
- Estimates of AFQT from PAY 97
- Are the current weights satisfactory?
- Design effect
- Equivalence of PAY 80 and PAY 97 AFQT means
- AFQT by subgroup.

Many of these topics touch on concerns expressed by the Defense Advisory Committee (DAC) on Military Personnel Testing or the Norming Advisory Group (NAG).

Background

- Our initial analysis was done on unweighted PAY 97 data
- This follow-on analysis is done on data weighted by NORC for the five psychometric edits
- We focus on two of these edits:
  - Edit 2: include language barrier cases and low response PSU. Outliers are deleted.
  - Edit 5: same as edit 2 except that highest grade completed is used in post-stratification weighting

Our initial analysis was conducted on unweighted data. This work is based on data weights developed by NORC according to specifications given by DMDC. DMDC gave specifications for 5 psychometric edits. In this report we focus on edit 2 and edit 5. The other edits represent minor variations.

- Edit 2: includes language barrier cases and low-response Primary Sampling Units (PSUs). Outliers are deleted.
- Edit 5: same as edit 2 except that the respondent's highest grade completed is used in post-stratification weighting.

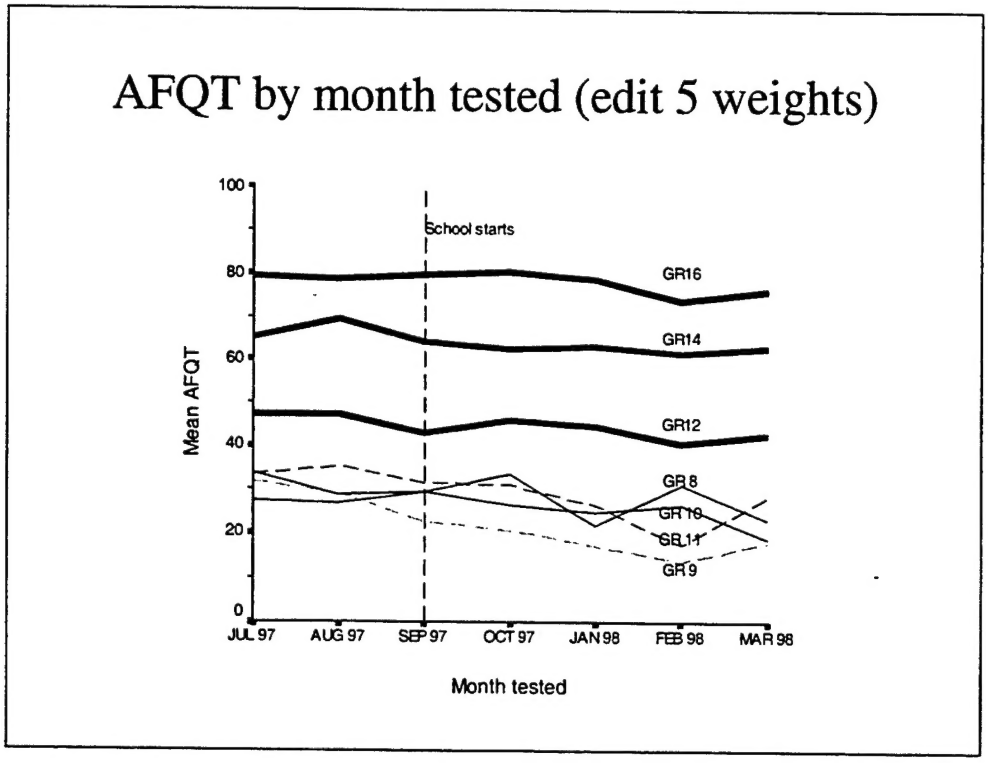
Weighting

- Edit 2: data are weighted to conform to CPS estimates of population by gender, ethnicity, and age.
- Edit 5: data are re-weighted to conform to CPS estimates of population by gender, ethnicity, and respondents education.
  - Neither age nor mother's education were used in this weighting

Weights for both edit 2 and edit 5 are lacking in some respects. We used one or the other in various stages in our analysis depending on where their relative strengths and weaknesses lie.

In edit 2 the data are weighted to conform to the Current Population Survey (CPS) estimates of population by gender, ethnicity, and age. Edit 2 weights are lacking in that they do not incorporate respondent's education or mother's education. However, for the variables on which they are focused, (age, gender, and ethnicity), the weights for edit 2 seem to have been correctly developed (unlike those for edit 5). Therefore, we used edit 2 weighted data as a starting point for our independent estimates of AFQT.

In edit 5 the data are weighted to conform to CPS estimates of population by gender, ethnicity, and respondent's education. Neither respondent's age nor mother's education were used in this weighting. Shortcomings in the development of weights for edit 5 will be seen to lead to distortions in the underlying population distributions by age and ethnicity. Despite these shortcomings, the edit 5 weights appear to lead to mean AFQT values that most closely approximate what we believe to be the correct value. We made use of edit 5 weights in those instances when we needed the most accurate representations of AFQT scores.



It had been expected that testing would begin in June 1997 and be completed before the fall school term began. Unfortunately, the testing was extended through April 1998 in order to approach the target sample sizes.

There has been some concern that persons tested after the start of a new fall term would score higher than otherwise due to additional schooling.

This chart shows mean AFQT by month tested by grade as of June 1997. This chart suggests that extending the testing into the next school year did not have any effect on mean AFQT scores. It is of course possible that two competing effects canceled each other. Persons tested later in the year might reasonably be expected to have scored higher as a result of additional learning. Conversely, persons tested later in the year might be expected to have been less willing to participate and hence produced lower scores.

Estimates of AFQT from external benchmarks

We did make estimates of AFQT from external benchmarks.

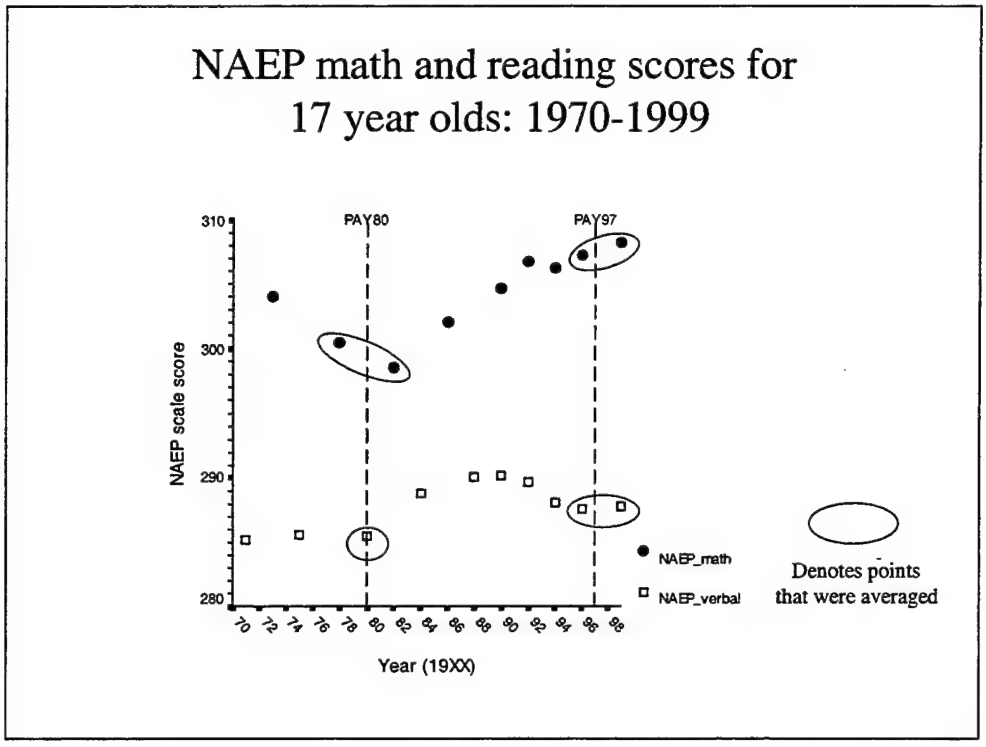
External benchmarks

- NAEP scale scores
  - 17 year olds
  - Math and reading
- Compared with PAY 80 and PAY 97
  - 18 year olds
  - AR, MK, and VE

We used the scale score data from the National Assessment of Educational Progress (NAEP).[2] The data cover 17-year-old youth tested in the spring of various years on math and reading skills. The math and verbal scale scores on the NAEP are known to be highly correlated to ASVAB.

We compared NAEP scores with ASVAB math and verbal scores on 18-year-old youth from PAY 80 and PAY 97.

2. NCES, *NAEP 1999 Trends in Academic Progress*, 2000.



This chart shows math and verbal scale scores for 17-year-old youth from 1970 through 1999. The chart also shows years when PAY (ASVAB) data was collected. In most cases, the years of NAEP testing did not correspond to years of PAY testing; therefore, we averaged the NAEP data from years that bracket the PAY years as indicated on the slide.

NAEP and ASVAB show similar changes

					Change: 1980 to 1997	
Category	Test	Age	"1997"	"1980"	Points	Std.dev.units
Math	NAEP math	17	307.7	299.5	8.20	.115
	ASVAB AR	18	48.89	48.77	0.12	.012
	ASVAB MK	18	52.44	49.82	2.62	.260
	ASVAB (MK+AR)/2	18	N/A	N/A	N/A	.137
Verbal	NAEP reading	17	287.7	285.5	2.20	.029
	ASVAB VE	18	48.90	48.50	0.40	.040

Source: *NAEP 1999 Trends in Academic Progress*, NCES
Inferred NAEP verbal std dev =75.2, math std dev = 71.4

This chart shows mean NAEP and PAY (ASVAB) scores for the "1980" and "1997" testing for youth of comparable ages. We used the PAY 97 edit 5 weights for this subset of the analysis because, as seen in later sections, they lead to AFQT estimates that we believe to be closest to the correct number. The average increase in NAEP math scores was 0.115 standard deviation; the average change in ASVAB math scores was 0.137 standard deviation. The average change in NAEP verbal scores was 0.029 standard deviation, and that for ASVAB verbal was 0.040. All ASVAB scores are on the 1980 score scale.

These changes in scores seemed to us to be consistent between the two tests and permitted us to use changes in NAEP to estimate expected mean AFQT scores for PAY 97.

Estimated changes in scores:1980-1997

						1980 to 1997 change in:			
	Mean "1997" scale score	Cases	Standard Error in mean	Inferred std dev.	Mean "1980" scale score	NAEP points	NAEP std. dev. units	Expressed in AFQT std. dev. units	Inferred AFQT points
NAEP Math	307.7	3,539	1.2	71.4	299.5	+8.2	.115	.115	
NAEP Reading	287.7	4,669	1.1	75.2	285.5	+2.2	.029	.029	
Average								.072	1.8[a]

Source: *NAEP 1999 Trends in Academic Progress*, NCES

a. Estimated via NAEP to AFQT_sum of standard scores to AFQT Percentile score. The result is 2.1 if going directly from NAEP to AFQT percentile score.

This chart shows how we estimated the change in ASVAB scores from 1980 to 1997.

We calculated the change in NAEP math and verbal scores, expressed them in standard deviation units, assumed that changes in ASVAB would be the same in terms of standard deviation units, and averaged the change in standard deviation units. We found that the average change in NAEP math and verbal scores was 0.072 standard deviation units.

We then converted this change into a change in AFQT expressed in standard score units and then converted that into a change in AFQT percentile units. The final estimate was an increase of 1.8 AFQT percentile units.

Mean AFQT in PAY 1997 estimated from changes in NAEP

Nominal PAY 1980 mean AFQT	50.0
Plus change inferred from NAEP:	+1.8
Expected PAY 1997 mean AFQT:	51.8

The nominal mean AFQT in PAY 80 was 50.0. If we then add the expected increase of 1.8 from the period 1980 to 1997, we get an estimate of 51.8 for mean AFQT in PAY 97.

Mother's education

In our earlier report on this data (CAB 99-66), we identified mother's education (a common proxy for social-economic status)as an important correlate of AFQT score. In this section we discuss the availability of data on mother's education and whether it is necessary to weight the data by this variable.

Update on mother's education data

- Mother's education was not collected for everyone in the ETP sample
- As of our last report (CAB 99-66), it was available for only 50% of the sample
- Recently, DMDC has obtained additional data from BLS
- Currently for the ETP edit 5 sample (6,143 cases)
  - 3,002 have mother's education from original file
  - 1,120 updated from BLS
  - 2,021 still missing mother's education (33%)

Unfortunately, mother's education was not collected for everyone in the sample. Originally it was missing for half of the sample. Now, after integrating additional data from BLS, we were still missing mother's education for about 1/3 of the sample.

Logistic regression on having vs not having mother's education data

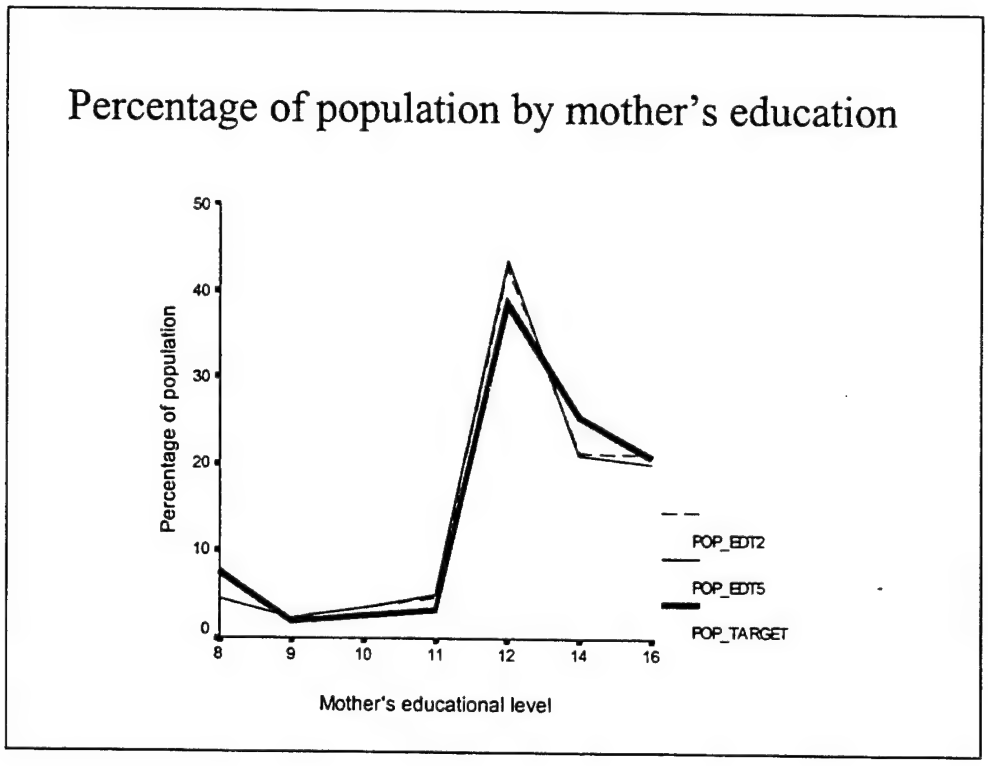
Variable	Significance level
AFQT	0.0000
Black	0.4669
Age	0.0000
Higrade	0.0968
Gender	0.0009
Hispanic	0.2117
White	0.0015
Cross sectional data	0.7634
Constant	0.0000

Red indicates a high degree of confidence that the variable differs between the group with mother's education data and the group without it.

This slide summarizes an analysis to determine whether the sub-samples with and without mother's education differ with respect to important variables. We defined a dummy variable (0/1) depending on whether mother's education was available for that record. We then did a logistic regression on the important variables to see whether any of them were important predictors of the dummy variable.

The results show that AFQT, age, gender, and race are significant predictors of the dummy variable. Hence, the sub-samples with and without mother's education differ in the underlying variables—i.e., AFQT and its correlates.

We were missing mother's education for a non-random 1/3 of the sample.



This chart shows the distribution in mother's education for the target population and the PAY 97 sample using
weights 2 and 5.

The PAY 97 sample appears to be missing small numbers of persons from both the high and low ends of the spectrum. There is an excess of cases with mother's education of level 12.

Does it matter if we don't have mother's education?

Not in aggregate, not for measures of central tendency

Mother's education level	Edit 2 weighted mean AFQT	PAY 97 edit 2 weighted population	Target population distribution
8	27.8	4.5	7.5
9	34.8	2.3	1.9
10	32.3	3.5	2.6
11	38.3	4.6	3.2
12	52.3	42.7	38.6
14	63.4	21.3	25.4
16	72.2	21.1	20.8
Total		100.0	100.0
Mean AFQT		56.0 ≈	56.1

In this slide we attempt to answer the question: Does it matter that we don't have mother's education to use in a weighting scheme?

We took the edit 2 data and estimated the mean AFQT with the knowledge that race, gender, and age have been properly taken into account by weight 2. We listed the mean AFQT for each interval of mother's education. We also listed the distribution by mother's education of the sub-sample having mother's education and for the target population.

We then estimated the mean AFQT with edit 2 weights for the sub-sample having mother's education by multiplying the mean AFQT per interval times the fraction of the population that is assumed in that interval and summing over all intervals.

AFQT = (27.8)*(.045) + (34.8)*(.023) + (32.3)*(.035)+(38.3)*(.046)+

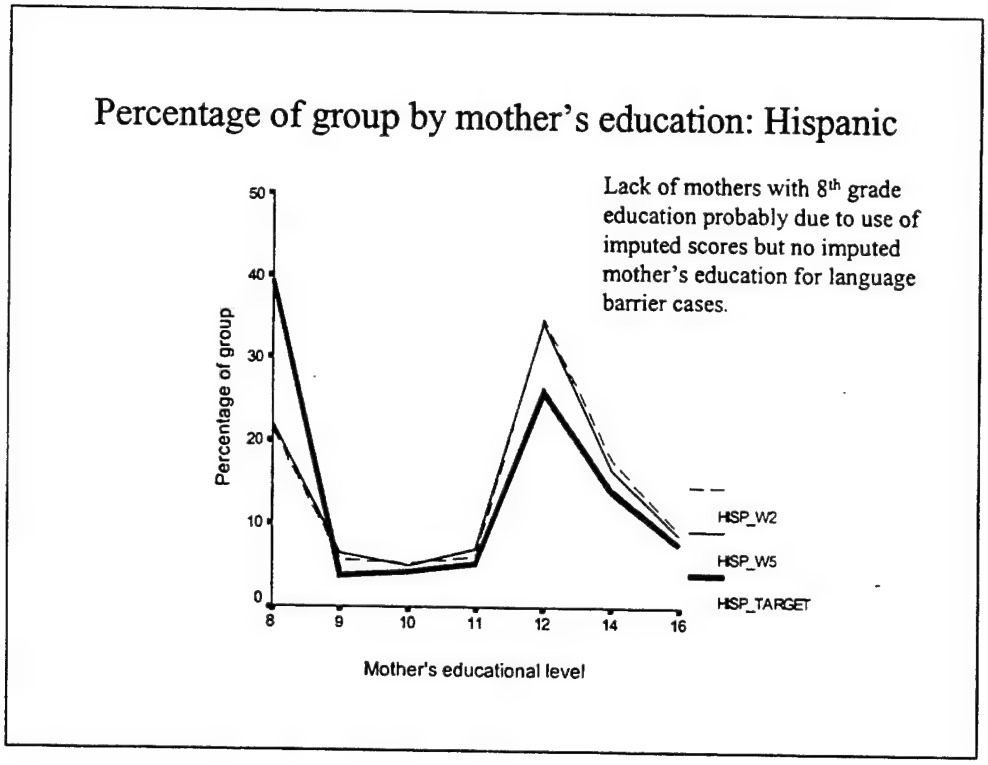
(52.3)*(.427)+(63.4)*(.213)+(72.2)*(.211) = 56.0.

Similarly, we estimated the mean AFQT if the data were weighted to the target population.

AFQT = (27.8)*(.075) + (34.8)*(.019) + (32.3)*(.026)+(38.3)*(.032)+

(52.3)*(.386)+(63.4)*(.254)+(72.2)*(.208) = 56.1.

Both outcomes were similar. Hence, no large errors were likely to be made in measures of central tendency if we did not weight on mother's education. Larger errors may possibly be made in distributional statistics.



This chart shows the population distributions in mother's education for Hispanics. At first glance this looks problematic. We appeared to be missing a large number of cases with very low levels of mother's education. However, upon inspection, this effect was associated with the large number of language barrier cases that have imputed AFQT scores but no imputed mother's education levels. This imputation is beyond the scope of our analysis. Once this imputation has been done, we expect that the weighted sample distribution will closely match that of the target.

Estimates of AFQT from PAY 97

We now turn to estimating the AFQT from the PAY 97 data. Because none of the current weights were satisfactory, we made a special estimate.

We started with the weighted data from edit 2. It correctly weights the data by gender, age, and ethnicity. We then modified the result to incorporate respondent's education. We called the result edit 2 mod 1.

Estimated impact of correct weighting
(using edit 2, and correcting for higrade: edit 2 mod1)

Respondent's education	Edit 2 weighted Mean AFQT	PAY 97 edit 2 weighted distribution	Target population distribution
8	27.4	2.0	2.5
9	22.3	1.7	2.0
10	25.8	3.8	3.5
11	27.5	8.0	7.9
12	43.4	32.1	36.1
14	63.7	40.7	41.2
16	78.0	11.8	6.8
Total		100.0	100.0
Mean AFQT		53.1 ≠	51.4

As before, we estimated the mean AFQT by multiplying mean AFQT in each educational interval by the edit 2 weighted population and by the target population. The mean weighted by the target population is the answer we sought.

We estimated the mean AFQT from edit 2 weighted population as:

AFQT = (27.4)*(.020) + (22.3)*(.017) + (25.8)*(.038)+(27.5)*(.080)+

(43.4)*(.321)+(63.7)*(.407)+(78.0)*(.118) = 53.1.

We estimated the mean AFQT from the target population distribution as:

AFQT = (27.4)*(.025) + (22.3)*(.020) + (25.8)*(.035)+(27.5)*(.079)+

(43.4)*(.361)+(63.7)*(.412)+(78.0)*(.068) = 51.4.

The estimate of 51.4 represents our best estimate of mean AFQT if the sample had been correctly weighted on gender, age, ethnicity, and respondent's highest grade completed.

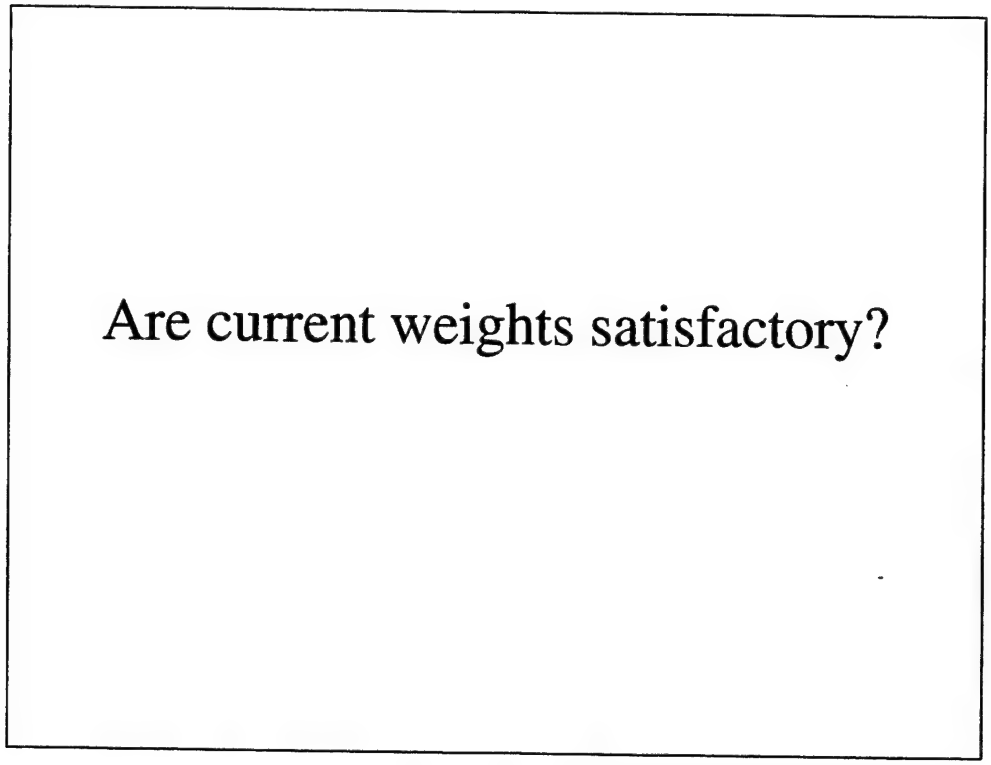
Mean PAY 97 AFQT by edit options

Options	Mean AFQT
Unweighted CX sample	54.7
Edit 1	53.0
Edit 2	53.1
Edit 3	53.1
Edit 4	52.6
Edit 5	52.0
Edit 2: mod1, (our best estimate from PAY 97)	51.4
Estimate from NAEP	51.8

This chart shows a summary of mean AFQT from various data edits (weights). The mean AFQT from the unweighted cross-sectional data is 54.7. As the sample is further refined, the mean AFQT is reduced.

Our best estimate of mean AFQT (edit 2 mod 1) is 51.4. This differs little from our estimate from NAPE of 51.8.

As previously noted, although the edit 5 weights are flawed and are based on distorted population distributions, they do lead to estimates of mean AFQT that are closest to what we believe to be the correct value.



We now turn to whether the current weights are satisfactory. The short answer is no.

Problematic population distributions age and ethnicity (weighted edit 5)

		Percentage of total group		
Variable	Value	Edit 5	Target	Edit 5-target
Age	18	20.8	17.3	3.5
	19	18.2	17.8	0.4
	20	17.6	17.0	0.6
	21	16.8	16.5	0.3
	22	13.9	15.4	-1.5
	23	12.7	16.0	-3.3
		100.0	100.0	
Ethnicity	White	72.1	70.6	1.5
	Black	14.5	14.8	-0.3
	Hispanic	13.4	14.6	-1.2
		100.0	100.0	

This chart illustrates the problem with the most recent of the current weights, edit 5. Recall that edit 5 weights only on gender, ethnicity, and respondent's educational level. It lets the age variable "float." It also combines some ethnicity cells with small cell populations. As a result, weighted data from edit 5 does not match the target population (which is essentially the Current Population Survey). In contrast, edit 2 weights lead to population distributions (see appendix) that closely match the CPS targets.

In our opinion, the lack of match between the edit 5 population distributions and the target is unacceptable.

Current weights are unsatisfactory

- W2 does not include either respondent's or mother's education
- The most refined weight we have is W5
  - But it does not include age or mother's education
- As a result the current weights lead to:
  - Incorrect estimates of AFQT
  - Incorrect underlying population distributions
- Given the scrutiny that this data will get from social scientists, these distortions are unsatisfactory, and, hence, the weights are unsatisfactory.

This slide summarizes the situation with respect the viability of current weights for PAY 97.

Weight 2 (edit 2) is not satisfactory because it does not include respondent's educational level.

Weight 5 (edit 5) includes respondent's educational level but distorts the age and ethnicity distributions. It is also unsatisfactory.

As a result the current weights lead to incorrect estimates of AFQT (and presumably other variables). Perhaps more importantly they distort the underlying population distributions.

This data set has the potential to be very useful in social science research. Given the scrutiny that the data will get from social scientists, these distortions are unsatisfactory, and, hence, the weights are unsatisfactory.

The weights must be corrected.

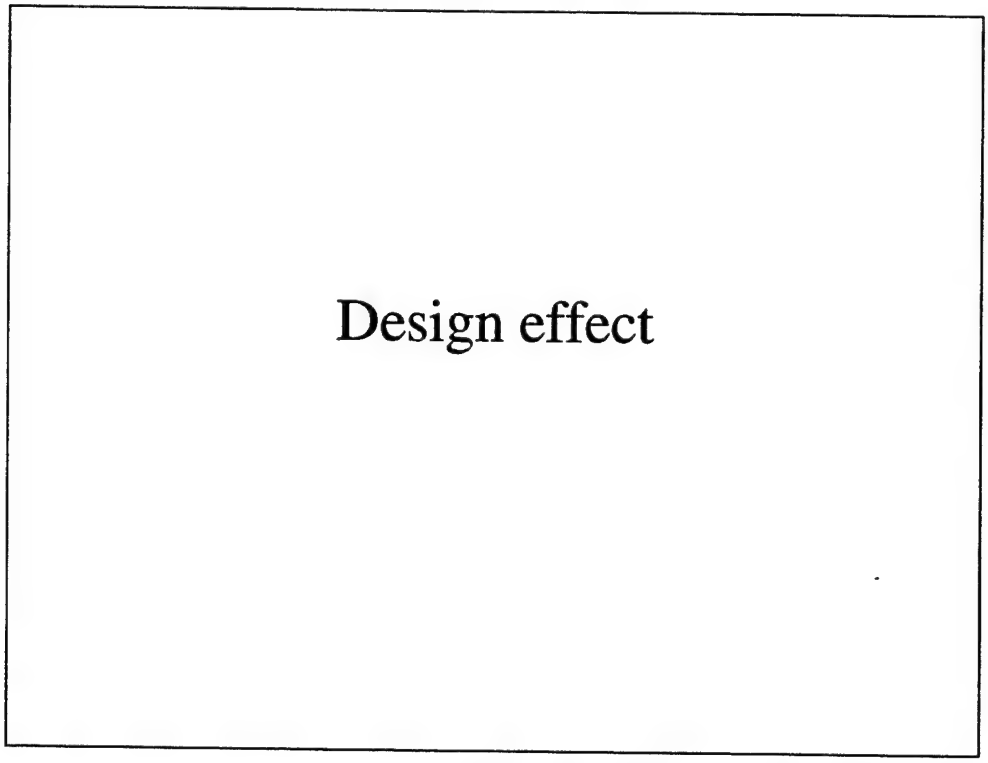
Is it possible to develop satisfactory weights for PAY 97?

- NORC initially weighted data on 3 variables:
  - Race, gender, and age
- CNA recommended weights on 5 variables:
  - Race, gender, age, education, and mother's education
- NORC re-weighted data on a different 3 variables:
  - Race, gender, and education
  - Not because they are perverse, but because of sample size
- We are missing mother's education for a non-random 1/3 of our sample
  - But this variable doesn't seem to be important in aggregate
- Many cells are small and may limit our ability to fix problems by weights.

Given that the weights must be corrected, the next question is whether it can be done. The short answer is probably yes. The sample must be weighted on ethnicity, gender, age, and respondent's education.

It appears that it would be impractical to weight on mother's education. Fortunately, it appears that this omission will not introduce much error.

It will be a challenge to calculate weights on the four variables noted above given the reality of small cell populations. We believe this to be the reason that NORC has always calculated weights on three (never four) variables. We strongly recommend against combining small cells across variables as was done by NORC in edit 5.



We now move to a discussion of the "design effect."

What is the design effect?

- It is a factor that expresses the inefficiency of a sample relative to a simple random sample:
  - Clustering reduces sampling efficiency
  - Oversampling reduces sampling efficiency
  - Stratification increases sampling efficiency
- Effective sample size is estimated as:
  - Actual sample size /design effect
- Why do we need to know it?
  - We need it to estimate statistical errors in PAY 97
- When will we know it?
  - When DMDC finalizes the sample and NORC calculates it
  - Until then we must use an estimate

The design effect is a factor that expresses the inefficiency of a sample relative to a simple random sample. Clustering and oversampling both reduce sampling efficiency. On the other hand, stratification increases sampling efficiency. All three procedures were used in PAY 97.

The effective sample size is the actual sample size divided by the design effect. We needed to know the design effect in order to estimate statistical errors in PAY 97.

The calculation of the design effect is a complex procedure and is normally left until the sample weights have been finalized. Once DMDC has done this, NORC would presumably calculate the design effect. Until then we must use an estimate.

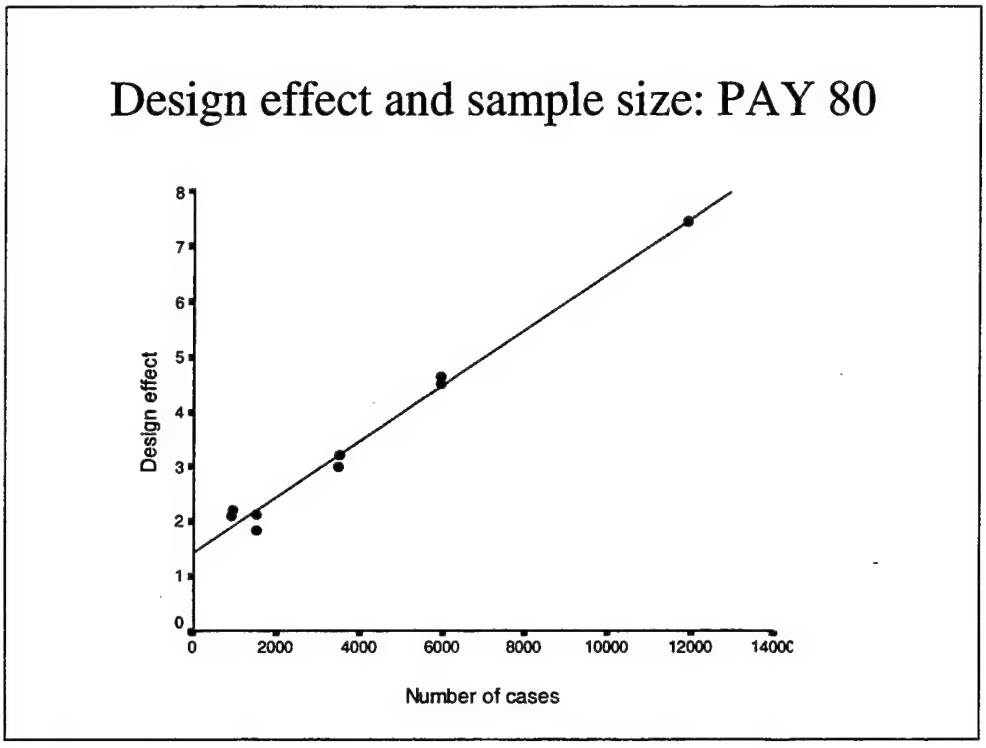
The design effect in PAY 80[a]

Gender	Race / ethnic	Number of cases	Design effect
Male	White	3,544	3.2164
	Black	1,517	1.8253
	Hispanic	908	2.1018
	Sub-total	5,969	4.6307
Female	White	3,499	2.9946
	Black	1,511	2.1147
	Hispanic	935	2.2091
	Subtotal	5,945	4.5057
Total		11,914	7.4373

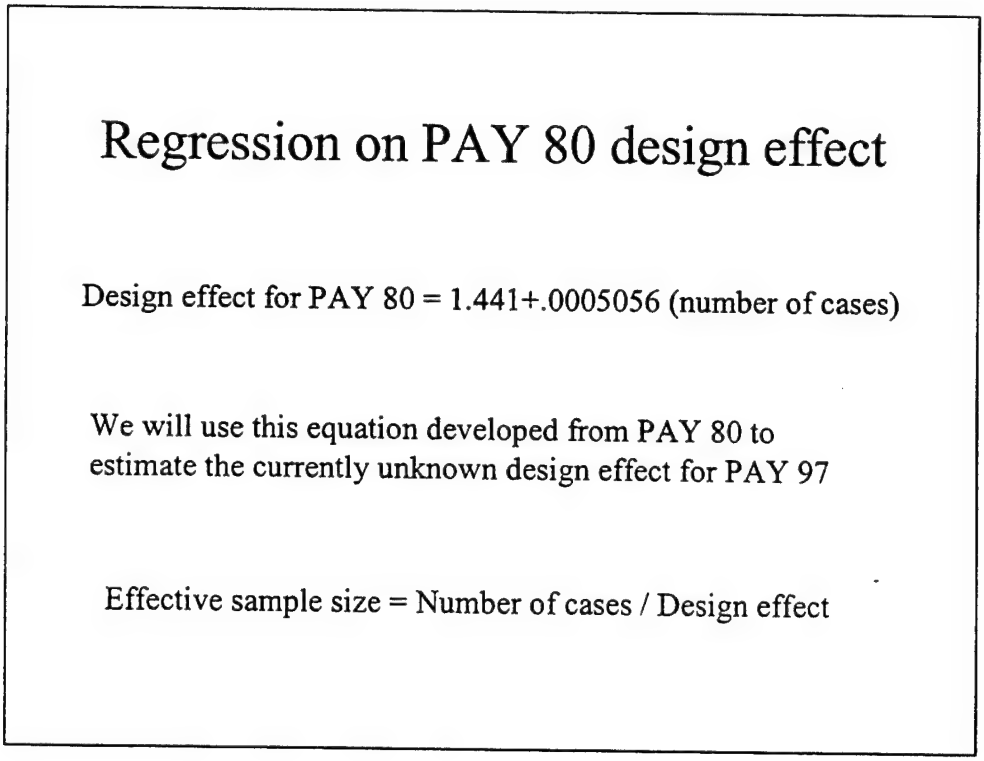
a. For mean AFQT scores as calculated by NORC.

This slide tabulates the design effect for mean AFQT scores as estimated by NORC for various subgroups in PAY 80. We assumed that these estimates are correct.

The clustering, oversampling, and stratification techniques used in PAY 97 are similar, although not identical, to those used in PAY 80. In the absence of better information, we worked to generalize the PAY 80 results so that we could apply them to the PAY 97 data.



On this slide we have plotted the design effect for PAY 80 versus the number of cases. The relationship is seen to be approximately linear with sample size. We have fitted a regression line through the data.



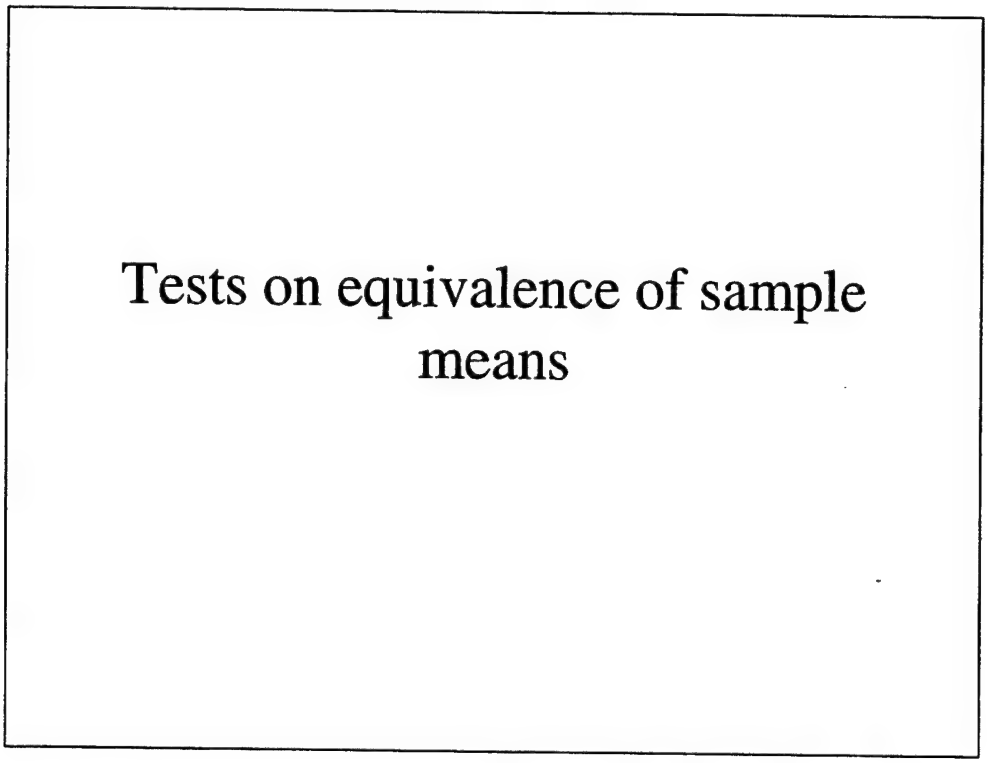
The regression equation for design effect as a function of the number of cases is shown on this slide. The data on which the equation is based are from PAY 80, but PAY 97 used the same sampling strategy, and the relationship is likely to provide a good estimate for the current data.

PAY 97 sample sizes

	Full sample			Sample with mother's education		
Group	Cases	Design effect[1]	Effective cases	Cases	Design effect[1]	Effective cases
White	3,402	3.16	1,076	2,501	2.71	924
Black	1,315	2.11	624	808	1.85	437
Hispanic	1,416	2.16	656	774	1.83	422
Male	2,788	2.85	978	1,796	2.35	765
Female	3,351	3.14	1,069	2,289	2.60	881
Total	6,134	4.45	1,351	4,085	3.51	1,165

1. Estimated from PAY 80.

Effective sample size is shown here for major subgroups, both for the full sample and for the sample for which we know mother's education. Note that the effective sample sizes are quite modest in many instances. For example, our total PAY 97 sample of 6,134 cases is equivalent to a simple random sample of only 1,351 cases.



We now discuss tests on the equivalence of sample means for AFQT scores.

AFQT from PAY 80

- Nominal mean AFQT is 50.0
- Actual mean AFQT is 50.4

The nominal mean AFQT from PAY 80 was 50.0, i.e., the score scale was designed to have a mean of 50. However, the actual mean is 50.4. We addressed the question of whether the PAY 97 results differ significantly from the PAY 80 results.

Effective sample size: PAY 80 and PAY 97

Sample	Sample size	Estimated design effect[a]	Effective sample size[b]
PAY 80:	9,173	6.08	1,509
PAY 97:	6,134	4.45	1,351

a. Design effect is a function of sample stratification and clustering, which was similar in both PAY 80 and PAY 97. The design effect for PAY 97 is estimated from that used in PAY 80.
b. Effective sample size is the size of a simple random sample of equivalent statistical power.

The design effect and effective sample size for the two samples is shown here.

Test for significance of difference between two means

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

We used a standard z-test for the significance of the difference between the means.

Tests for significance of differences in AFQT means: PAY 80 vs PAY 97

Comparison		Mean AFQT		Standard deviations		Effective sample size		Test statistic	Differences significant at:	
PAY 80	PAY 97	PAY 80	PAY 97	PAY 80	PAY 97	PAY 80	PAY 97	Z	90%	95%
Nominal AFQT	Edit 2	50.0	53.1	28.9	28.6	1,509	1,351	2.88	Yes	Yes
Nominal AFQT	Edit 5	50.0	52.0	28.9	28.3	1,509	1,351	1.86	Yes	No
Nominal AFQT	Edit 2.1	50.0	51.4	28.9	28.6	1,509	1,351	1.30	No	No
Actual AFQT	Edit 2	50.4	53.1	28.9	28.6	1,509	1,351	2.51	Yes	Yes
Actual AFQT	Edit 5	50.4	52.0	28.9	28.3	1,509	1,351	1.49	No	No
Actual AFQT	Edit 2.1	50.4	51.4	28.9	28.6	1,509	1,351	0.93	No	No

This slide summarizes the results of the calculation on the significance of differences in the means.

We calculated the two different mean values of AFQT from PAY 80 and for three different edits (weights) used in PAY 97.

The mean AFQT (from PAY 97) from edit 2 is seen to be significantly different from the mean AFQT from PAY 80. However, as previously discussed, we believe that edit 2 is seriously flawed.

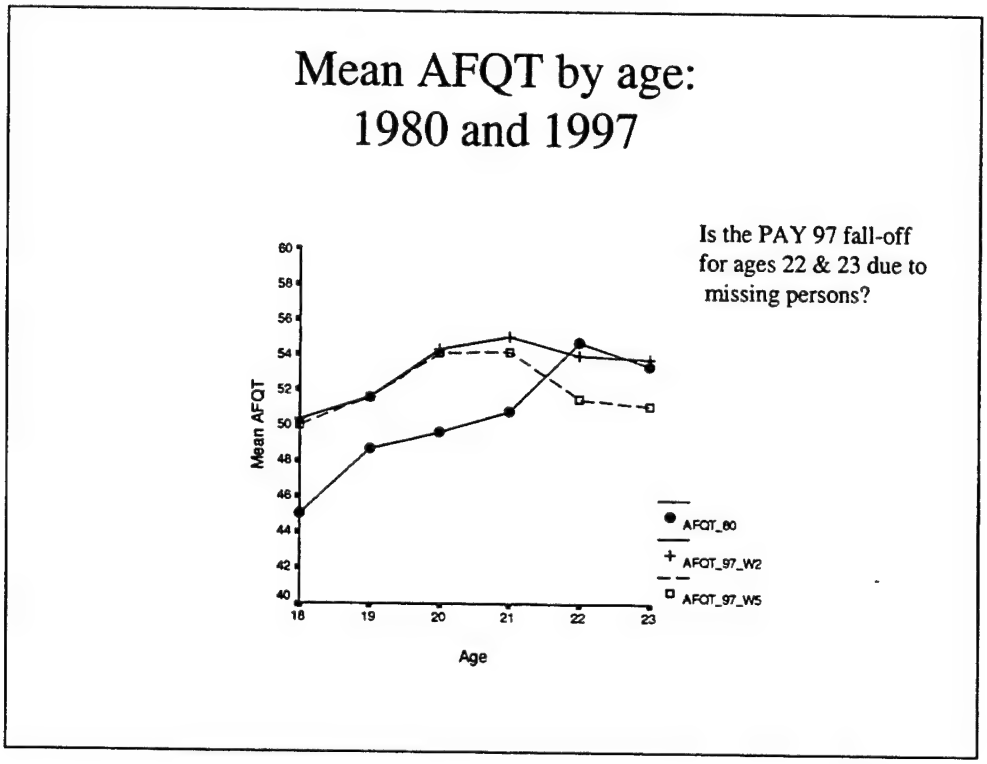
The mean AFQT (from PAY 97) from edit 5 is not significantly different from the mean AFQT from PAY 80 at the 95-percent level.

Our best estimate of the mean AFQT from PAY 97 is edit 2.1. It is not significantly different from the PAY 80 mean at either the 90-percent or 95-percent level.

AFQT by subgroup: 1980 and 1997

The aptitude test score data collected in PAY 97 has the potential to be of great value to policy-makers and social science researchers. Some of the more interesting results are shown in the following charts. Note that the data is weighted by weights 2 and 5—neither of which is satisfactory in our opinion. Hence, these charts should be considered only approximations to what the final correctly weighted results will show, and we have not shown any tests of significance on these data.

We also show these charts because they may be of use to persons attempting to make judgments on the credibility of the PAY 97 sample.



This chart compares the mean AFQT by age from PAY 80 with that for PAY 97 estimated using weights 2 and 5. All scores are expressed in the 1980 reference score scale.

We think that the downturn in mean AFQT for the PAY 97 for ages 22 and 23 is disturbing. This age group (22 and 23) represented the deepest part of the "hole" in the age distribution for PAY 97.

Mean AFQT by respondent's education: 1980 and 1997

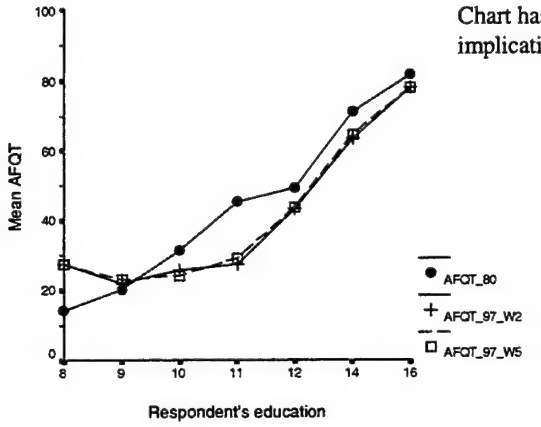
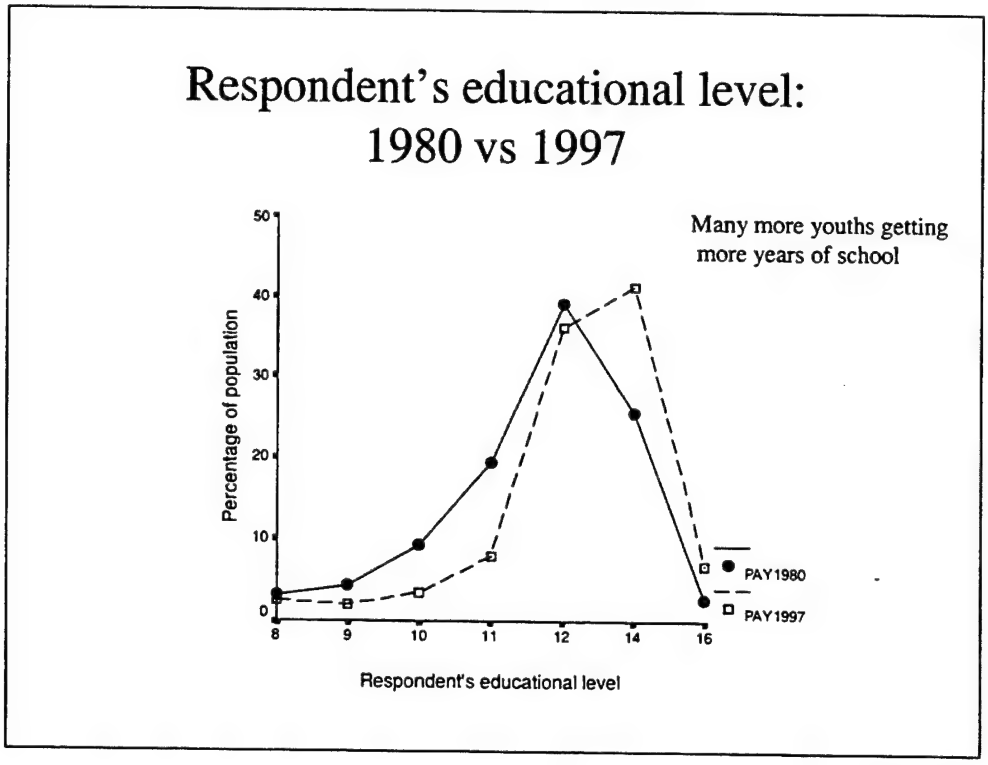


Chart has important policy implications, if true

This chart shows mean AFQT by highest grade completed as of June 1997 (ages 18 through 23).

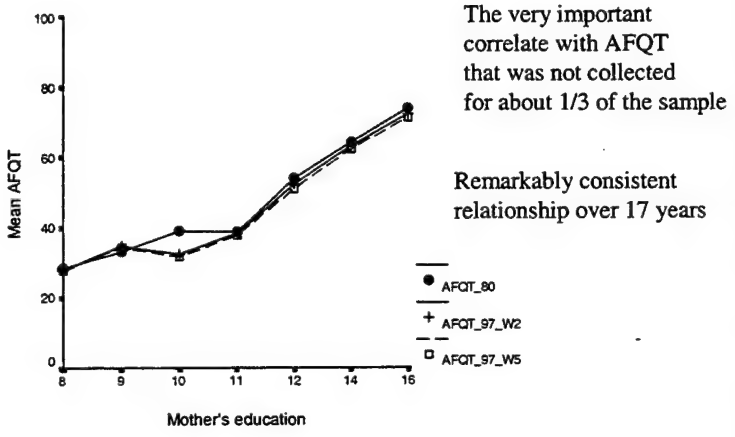
We see that PAY 97 persons generally score better than PAY 80 persons for the lowest levels of respondent's education. However, for all other levels, the PAY 97 sample scores below the PAY 80 sample.



This chart accompanies the previous chart to explain why the mean AFQT from PAY 97 is slightly higher than that for PAY 80, although at almost all levels of education the respondents from PAY 97 score worse.

The paradox may be understood from this chart, which shows that many more youths are going on to higher levels of education in PAY 97. One might say that youths are learning less at each grade but making up for it by going on to higher grades.

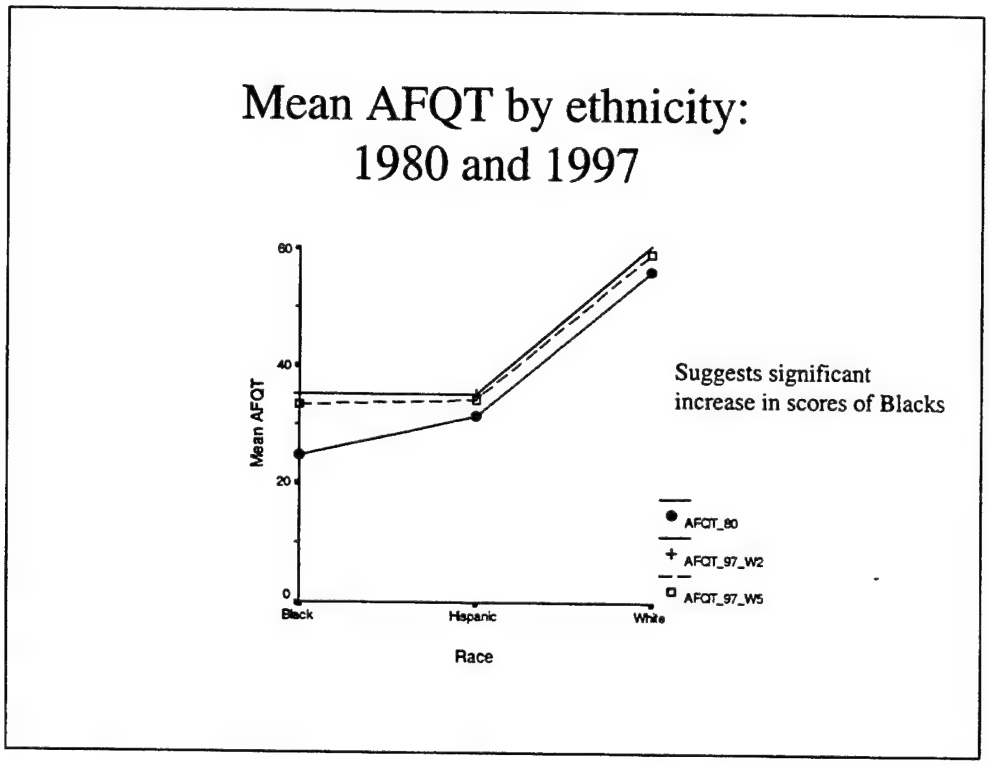
Mean AFQT by mother's education: 1980 and 1997



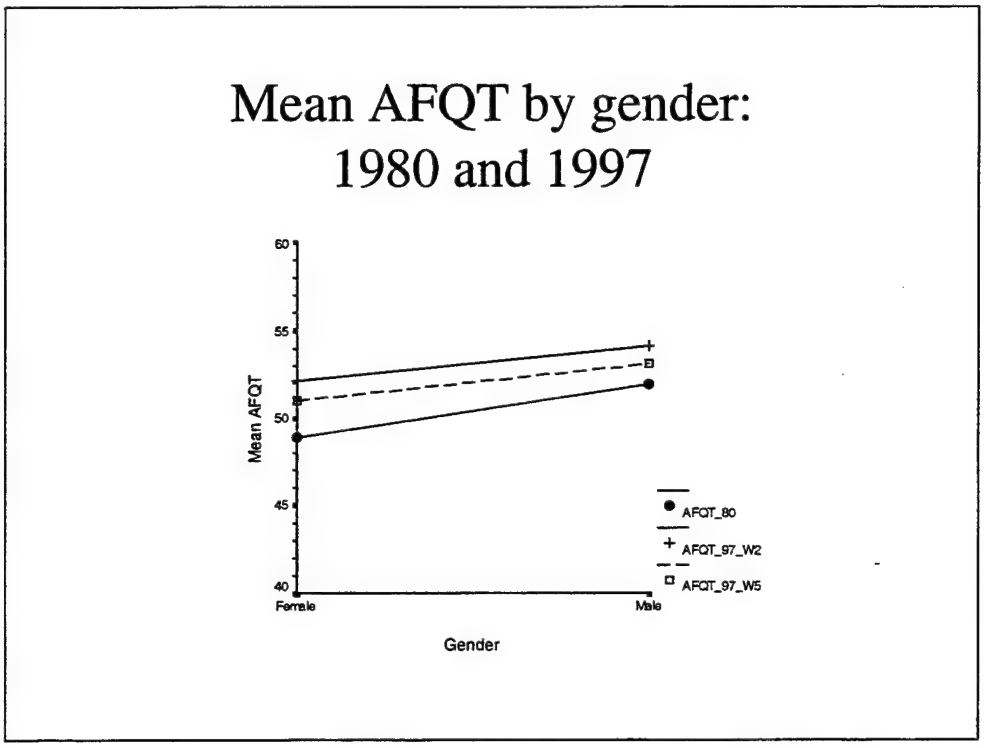
The very important correlate with AFQT that was not collected for about 1/3 of the sample

Remarkably consistent relationship over 17 years

This chart shows mean AFQT by mother's education. The relationship is remarkably stable over the 17-year period from 1980 to 1997.



This chart shows mean AFQT by race / ethnicity. Black youth appear to have made major strides in scores.



Both genders are seen to have made about the same small improvements from 1980 to 1997.

Conclusions

- External benchmarks lead us to expect about 51.8 for the mean AFQT.
- The likely mean AFQT is 51.4.
- New data on mother's education does not allow its use in weighting, which is not a problem in aggregate.
- There is no bias from the additional months of testing.
- NORC population control distributions are reasonable.

Our conclusions are summarized on this and the following two slides.

Conclusions (continued)

- Resulting weighted PAY 97 distributions are not satisfactory.
- They should be fixed.
- They probably can be fixed.
  - It appears that weighting on age, ethnicity, gender, and respondents education would be sufficient
    - This has not yet been done
  - Small cell populations are a serious problem

Conclusions (continued)

- Our best estimates of mean AFQT from the norming samples are:
  - PAY 97 51.4 $\pm$ 0.8
  - PAY 80 50.4 $\pm$ 0.7
- These two means are not statistically different at the traditional 90% or 95% confidence levels
- If this PAY 97 estimate should be confirmed, there would seem to be no compelling evidence that score scales for the enlisted testing program would need to be changed from those developed in PAY 80.
  - Particularly in light of the sampling difficulties observed in PAY 97.

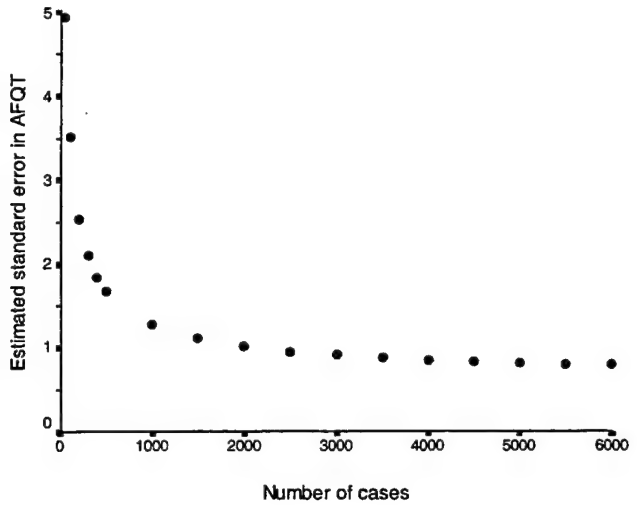
Appendix

Merging BLS file with new data on mother's education

- New BLS file has 7,367 records
  - 120 IDs have multiple records
- Editing rules for BLS file (7,247 cases after editing)
  - Delete all but first of exact duplicates
  - If duplicate IDs differ in mother's education, keep the one that agrees with mother's education on current SCF
  - Deleted 7 duplicates with unresolvable inconsistencies
- Merge new BLS with current SCF
  - Retain mother's education from original SCF
  - Supplement with values from BLS if available
- Result: PAY 97 edit 5 sample (6,143 cases)
  - 3,002 have mother's education from original file
  - 1,120 updated from BLS
  - 2,021 still missing mother's education (33%)

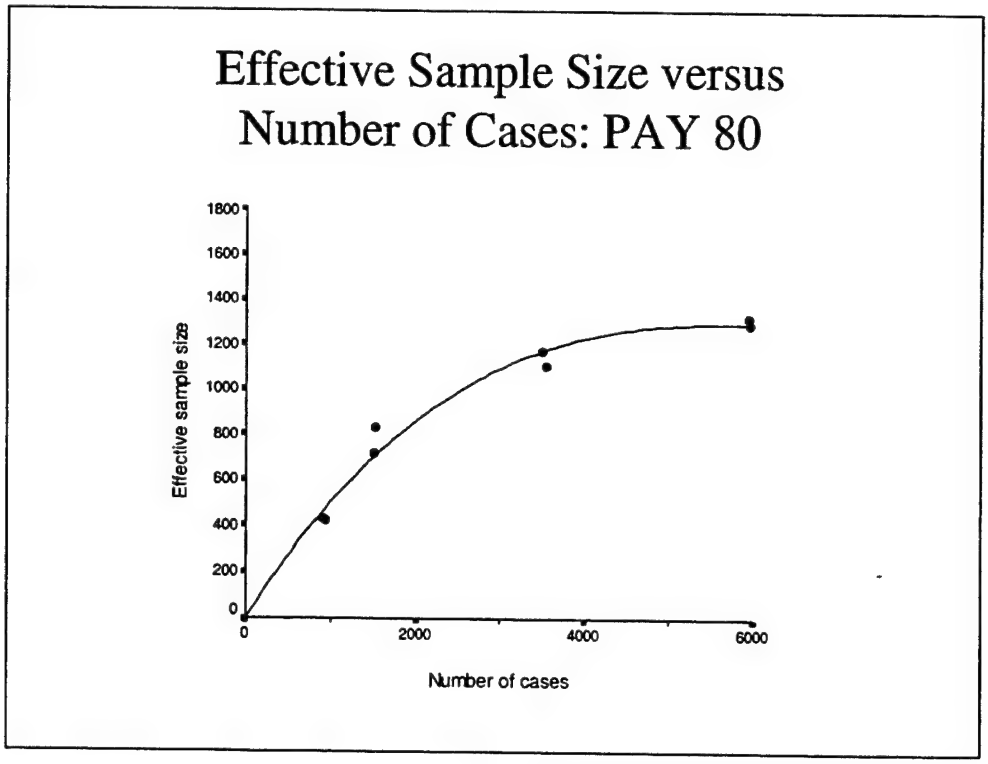
NOTES: SCF = Sample Control File.

Estimated Standard Error in Mean AFQT (PAY 97)



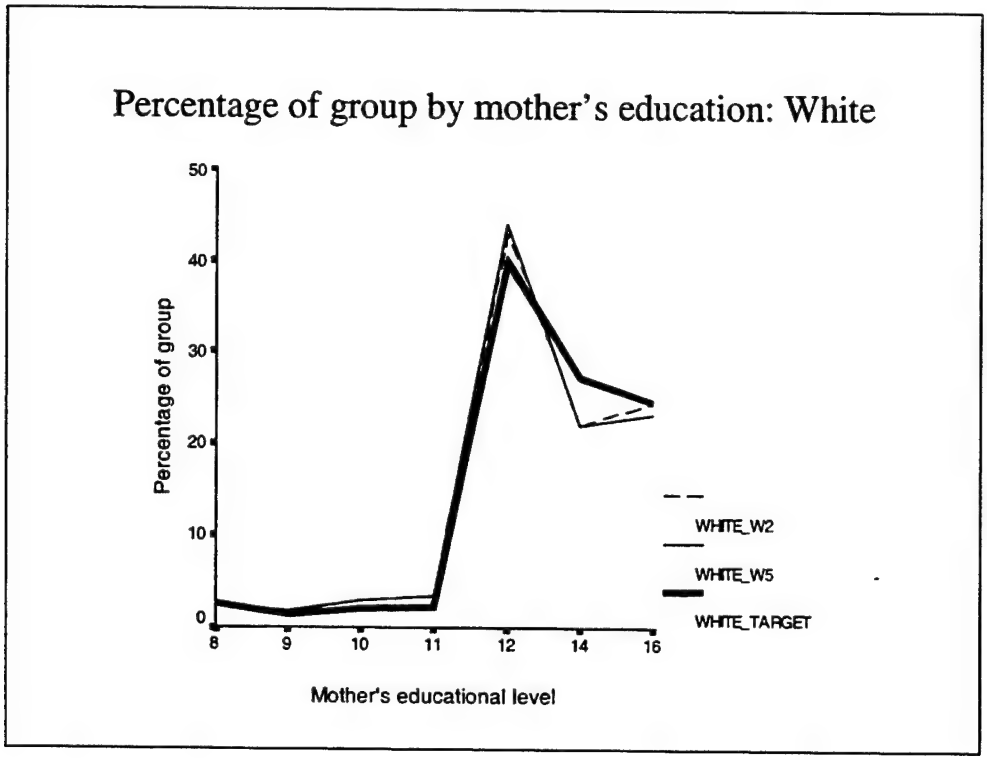
We can estimate standard errors in mean values as the standard deviation divided by the square root of the effective sample size. The effective sample size is the number of cases divided by the design effect.

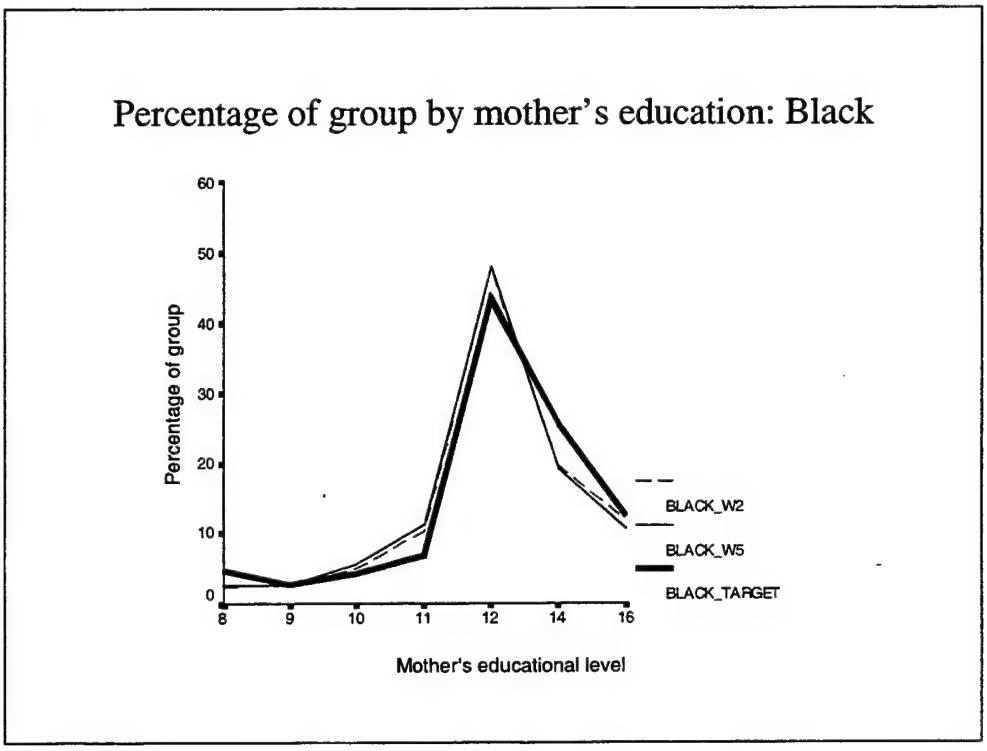
This slide shows estimates of standard error in mean AFQT for samples (or sub-samples) of various sizes.


Effective Sample Size versus Number of Cases: PAY 80
Effective sample size
1800
1600
1400
1200
1000
800
600
400
200
0
0
2000
4000
6000
Number of cases

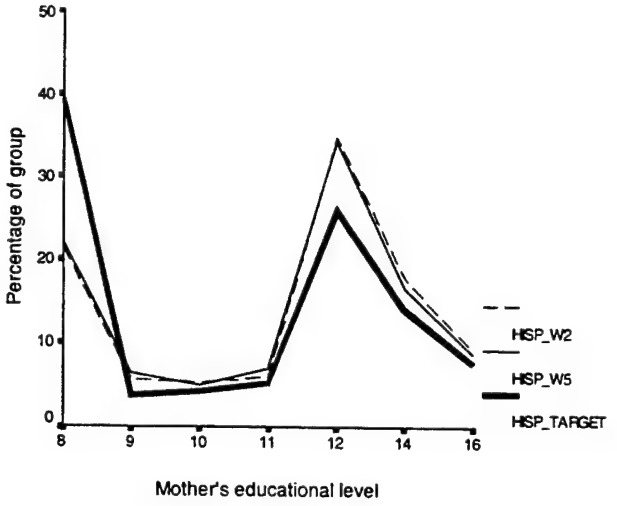
Population distributions

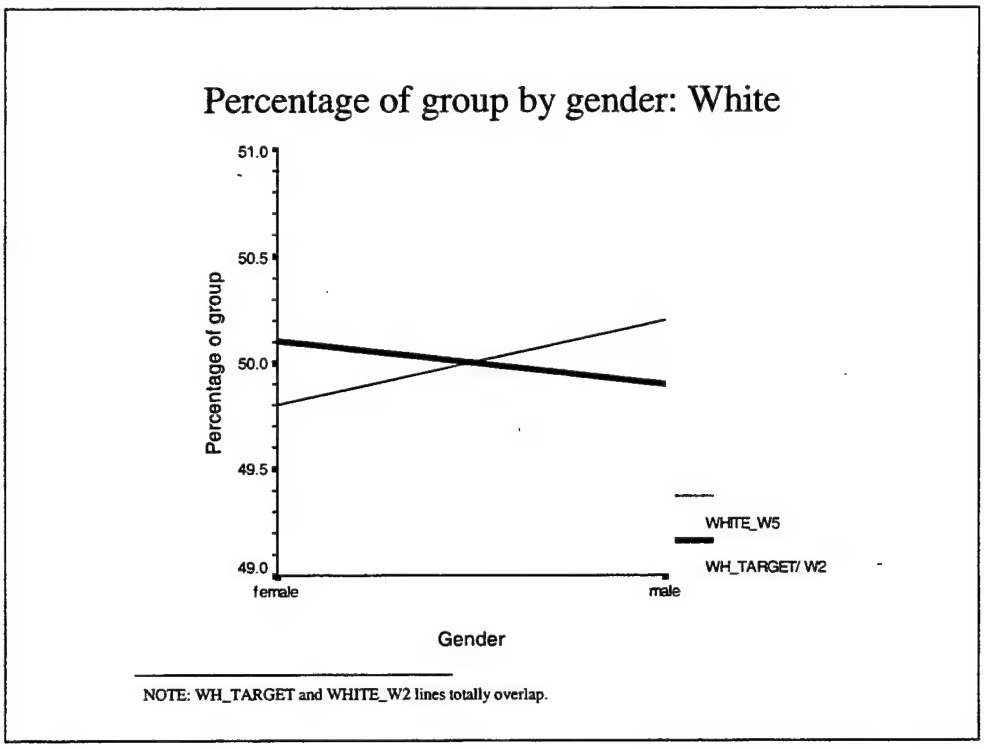
W2, W5, and target


Percentage of group by mother's education: White
Percentage of group
50
40
30
20
10
0
8
9
10
11
12
14
16
Mother's educational level
WHITE_W2
WHITE_W5
WHITE_TARGET


Percentage of group by mother's education: Black
Percentage of group
60
50
40
30
20
10
0
8
9
10
11
12
14
16
Mother's educational level
BLACK_W2
BLACK_W5
BLACK_TARGET

Percentage of group by mother's education: Hispanic



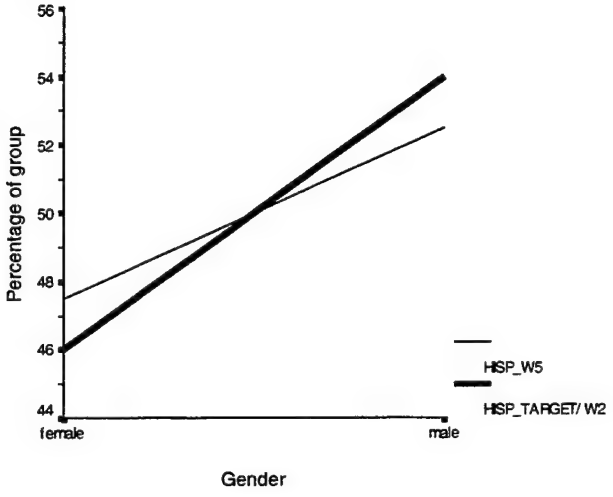


NOTE: WH_TARGET and WHITE_W2 lines totally overlap.

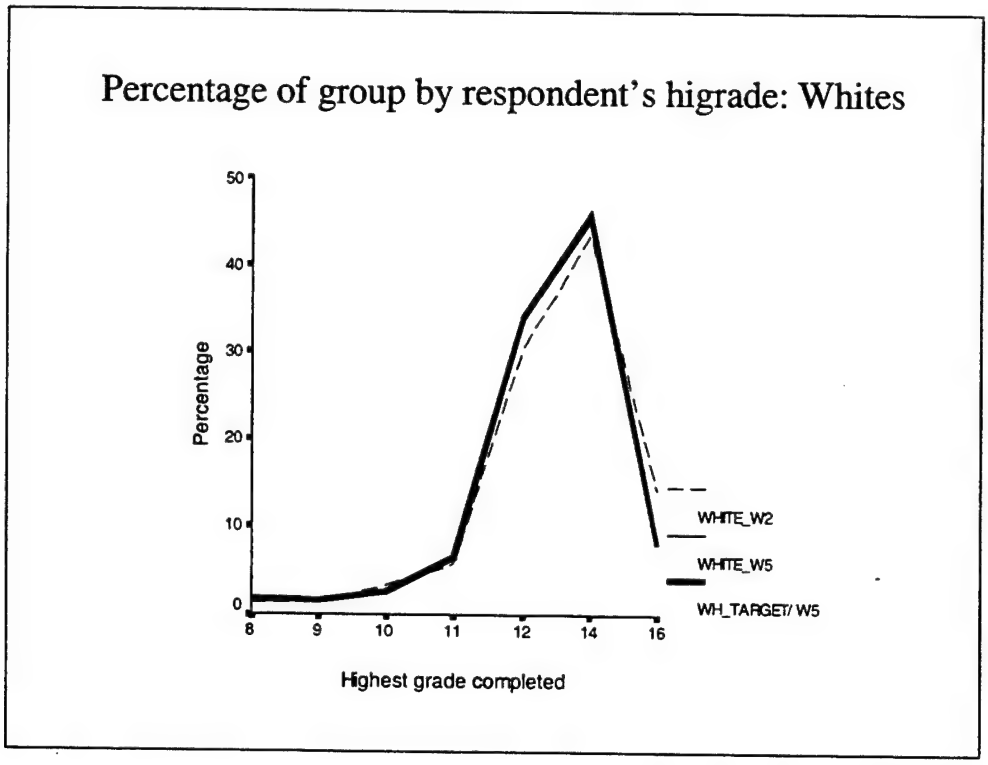
Percentage of group by gender: Black

NOTE: BL_TARGET and BL_W2 lines totally overlap.

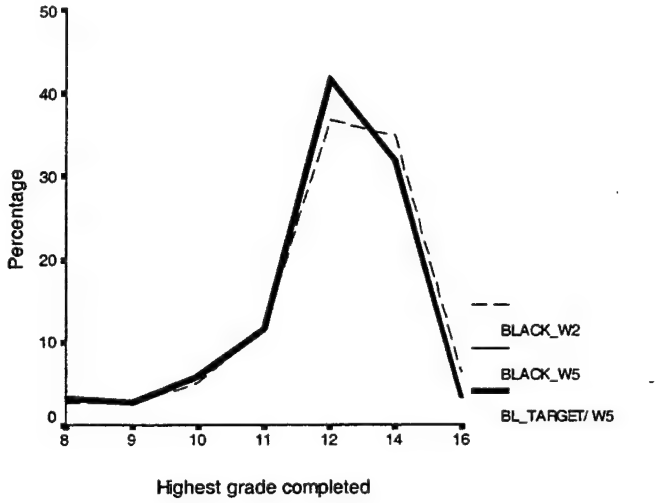
Percentage of group by gender: Hispanic



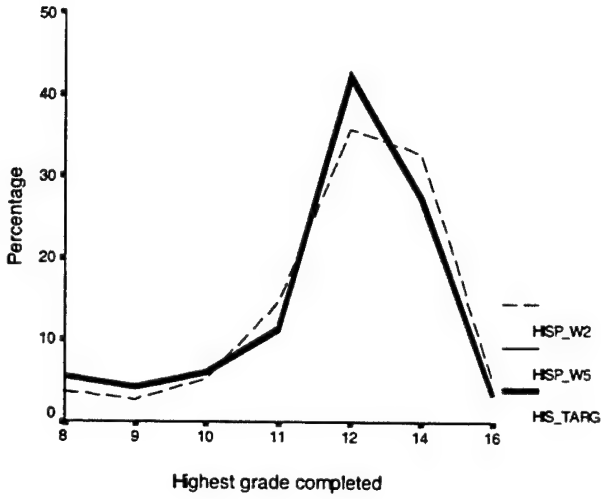
NOTE: HISP_TARGET and HISP_W2 lines totally overlap.


Percentage of group by respondent's higrade: Whites
Percentage
50
40
30
20
10
0
8
9
10
11
12
14
16
Highest grade completed
WHITE_W2
WHITE_W5
WH_TARGET/ W5

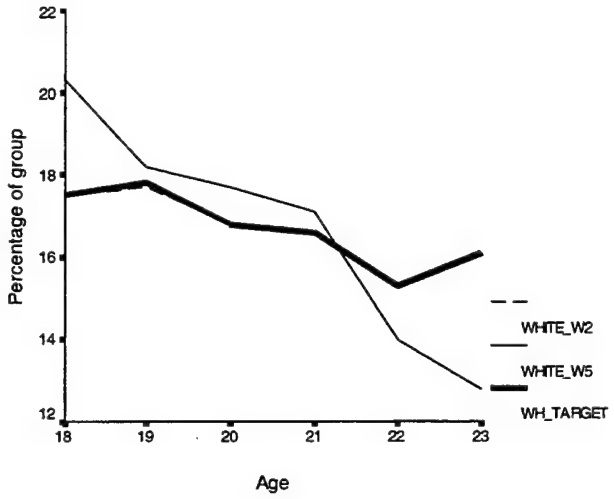
Percentage of group by respondent's higrade: Black



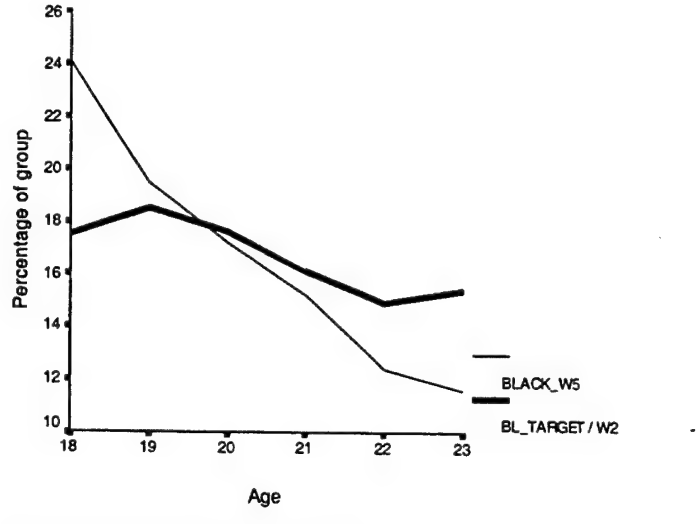
Percentage of group by respondent's higrade: Hispanic



Percentage of group by age: White

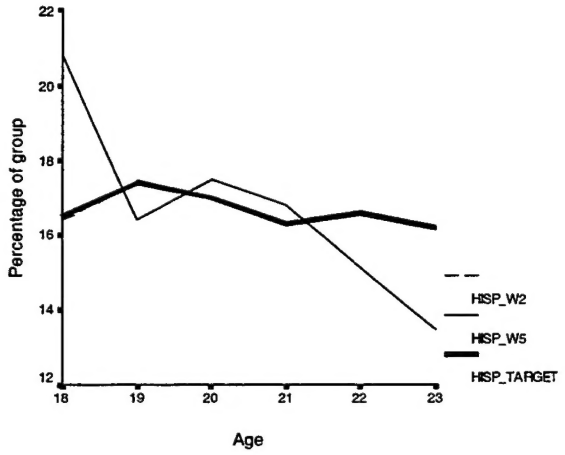


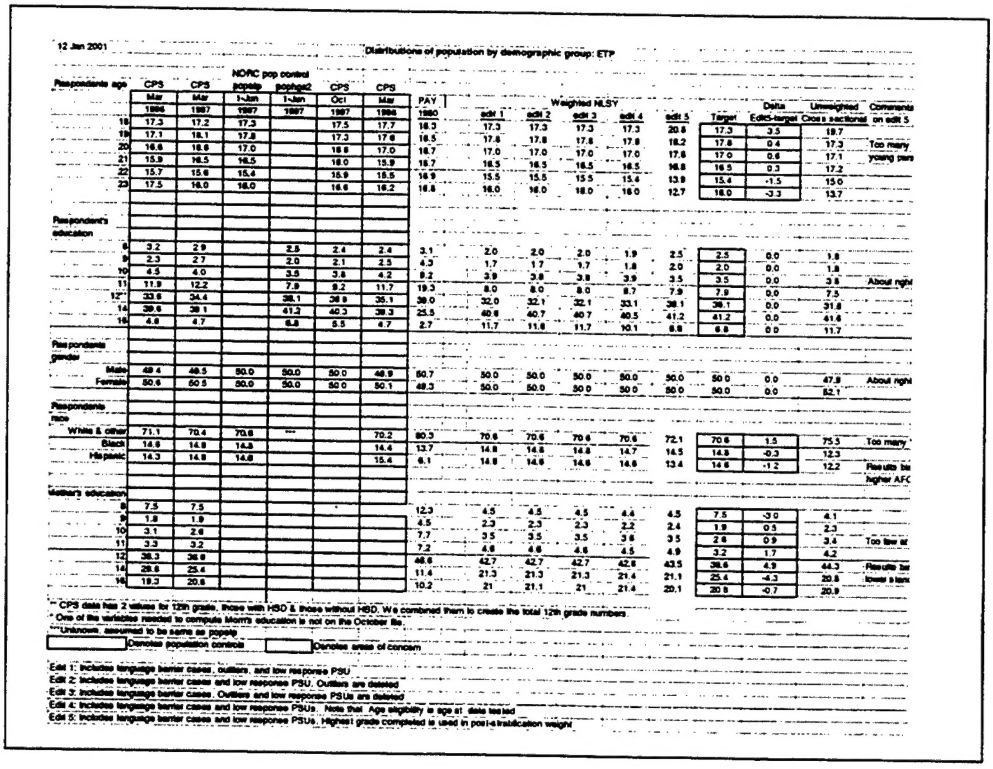
Percentage of group by age: Black



NOTE: BL_TARGET and BL_W2 lines totally overlap.

Percentage of group by age: Hispanic





12 Jan 2001

Distributions of population by demographic group: ETP

Respondents age	CPS Mar 1996	CPS Mar 1997	NORC pop control popsep 1-Jan 1997	popfeb2 1-Jan 1997	CPS Oct 1997	CPS Mar 1996	PAY 1990	Weighted NLSY edit 1	edit 2	edit 3	edit 4	edit 5	Target	Delta Edit5-target	Unweighted Cross sectional	Comments on edit 5
19	17.3	17.2	17.3		17.5	17.7	18.3	17.3	17.3	17.3	17.3	20.8	17.3	3.5	19.7	
19	17.1	18.1	17.8		17.3	17.8	18.5	17.8	17.8	17.8	17.8	18.2	17.8	0.4	17.3	Too many
20	16.8	16.9	17.0		16.8	17.0	16.7	17.0	17.0	17.0	17.0	17.6	17.0	0.6	17.1	young pare
21	15.9	16.5	16.5		16.0	15.9	16.7	16.5	16.5	16.5	16.5	16.8	16.5	0.3	17.2	
22	15.7	15.6	15.4		15.9	15.5	16.9	15.5	15.5	15.5	15.4	13.9	15.4	-1.5	15.0	
23	17.5	16.0	16.0		16.6	16.2	16.8	16.0	16.0	16.0	16.0	12.7	16.0	-3.3	13.7	
Respondent's education																
9	3.2	2.9		2.5	2.4	2.4	3.1	2.0	2.0	2.0	1.9	2.5	2.5	0.0	1.8	
9	2.3	2.7		2.0	2.1	2.5	4.2	1.7	1.7	1.7	1.8	2.0	2.0	0.0	1.8	
10	4.5	4.0		3.5	3.8	4.2	8.2	3.9	3.8	3.8	3.9	3.5	3.5	0.0	3.9	About right
11	11.9	12.2		7.9	9.2	11.7	19.3	8.0	8.0	8.0	8.7	7.9	7.9	0.0	7.5	
12*	33.6	34.4		38.1	38.9	35.1	38.0	32.0	32.1	32.1	33.1	38.1	38.1	0.0	31.8	
14	28.6	28.1		41.2	40.3	38.3	25.5	40.8	40.7	40.7	40.5	41.2	41.2	0.0	41.6	
16	4.8	4.7		6.8	5.5	4.7	2.7	11.7	11.8	11.7	10.1	6.8	6.8	0.0	11.7	
Respondents gender																
Male	49.4	46.5	50.0	50.0	50.0	48.9	50.7	50.0	50.0	50.0	50.0	50.0	50.0	0.0	47.9	About right
Female	50.6	50.5	50.0	50.0	50.0	50.1	49.3	50.0	50.0	50.0	50.0	50.0	50.0	0.0	52.1	
Respondents race																
White & other	71.1	70.4	70.8	---		70.2	80.3	70.6	70.6	70.6	70.6	72.1	70.6	1.5	75.3	Too many
Black	14.6	14.8	14.5			14.4	13.7	14.8	14.8	14.8	14.7	14.5	14.8	-0.3	12.3	
Hispanic	14.3	14.8	14.8			15.4	6.1	14.8	14.9	14.9	14.8	13.4	14.9	-1.2	12.2	Results bia
																higher AFC
Mother's education																
8	7.5	7.5					12.3	4.5	4.5	4.5	4.4	4.5	7.5	-3.0	4.1	
9	1.8	1.9					4.5	2.3	2.3	2.3	2.2	2.4	1.9	0.5	2.3	
10	3.1	2.6					7.7	3.5	3.5	3.5	3.8	3.5	2.6	0.9	3.4	Too few at
11	3.3	3.2					7.2	4.8	4.8	4.8	4.5	4.9	3.2	1.7	4.2	
12	38.3	36.8					48.8	42.7	42.7	42.7	42.8	43.5	38.6	4.9	44.3	Results bia
14	28.8	25.4					11.4	21.3	21.3	21.3	21.4	21.1	25.4	-4.3	20.8	lower a ten
16	19.3	20.8					10.2	21	21.1	21	21.4	20.1	20.8	-0.7	20.9	

* CPS data has 2 values for 12th grade, those with HSD & those without HSD. We combined them to create the total 12th grade numbers

+ One of the variables needed to compute Mom's education is not on the October file.

**Unknown, assumed to be same as popsep

Denotes population control Denotes areas of concern

Edit 1: Includes language barrier cases, outliers, and low response PSU

Edit 2: Includes language barrier cases and low response PSU. Outliers are deleted

Edit 3: Includes language barrier cases. Outliers and low response PSUs are deleted

Edit 4: Includes language barrier cases and low response PSUs. Note that Age eligibility is age at date tested

Edit 5: Includes language barrier cases and low response PSUs. Highest grade completed is used in post-stratification weight

12 Jan 2001

Distributions in AFQT by demographic group: ETP

Respondents age	Weighted NLSY 97												Weighted NLSY 80
	edit 1		edit 2		edit 3		edit 4		edit 5		unweighted CX		
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean
18	50.1	28.8	50.3	28.7	50.3	28.7	50.0	29.0	50.0	28.8	50.6	28.9	45.0
19	51.3	28.2	51.6	28.1	51.6	28.1	50.2	28.5	51.6	28.0	52.5	27.7	48.7
20	54.2	28.3	54.3	28.2	54.3	28.2	53.2	28.2	54.1	28.1	56.3	27.5	49.6
21	55.0	28.5	55.0	28.5	55.1	28.5	55.0	28.3	54.2	28.2	57.2	27.8	50.8
22	53.9	29.3	54.0	29.2	54.0	29.1	54.6	29.3	51.5	28.7	57.3	27.4	54.7
23	53.8	28.3	53.8	28.3	53.7	28.3	53.0	28.3	51.1	27.8	55.5	27.3	53.4
	53.0	28.6	53.1	28.6	53.1	28.5	52.6	28.7	52.0	28.3	54.7	27.9	50.0
Respondent's education													
8	26.9	21.2	27.4	21.2	27.4	21.2	28.0	20.3	27.4	21.3	29.3	21.1	14.4
9	22.2	17.9	22.3	17.9	22.5	17.9	22.1	17.7	23.1	18.2	21.2	16.1	20.3
10	25.6	17.9	25.8	17.9	25.8	17.9	25.8	18.0	24.3	18.0	27.4	17.2	31.3
11	27.4	21.0	27.5	21.0	27.5	20.9	28.3	21.8	29.1	21.4	30.8	20.7	45.3
12	43.3	25.4	43.4	25.4	43.4	25.4	44.2	26.0	43.7	25.5	44.9	25.0	49.4
14	63.6	24.6	63.7	24.5	63.7	24.5	63.9	24.5	64.7	24.2	64.5	24.1	71.2
16	78.0	18.9	78.0	18.9	78.0	18.9	78.2	18.9	78.2	18.7	78.6	18.1	81.8
Respondents gender													
Male	54.0	29.1	54.1	29.0	54.1	29.0	53.3	29.2	53.1	28.7	56.6	28.3	51.9
Female	52.1	28.1	52.1	28.1	52.1	28.0	51.9	28.1	51.0	27.9	53.0	27.5	48.9
Respondents race													
White & other	60.5	26.2	60.6	26.1	60.6	26.1	60.2	26.3	59.2	26.1	60.5	26.2	58.2
Black	35.0	24.3	35.1	24.3	35.0	24.2	34.2	24.0	33.3	23.6	33.2	24.4	24.6
Hispanic	35.1	27.6	35.1	27.6	35.2	27.6	34.8	27.6	34.2	27.1	41.6	25.9	31.3
Mother's education													
8	27.7	24.5	27.8	24.5	27.8	24.4	28.5	24.7	27.7	24.3	31.6	22.6	28.6
9	34.4	24.2	34.5	24.0	34.7	23.8	33.0	24.3	34.5	23.9	34.5	24.4	33.1
10	32.3	22.4	32.3	22.4	32.2	22.4	31.2	21.7	31.7	22.3	34.1	22.7	39.2
11	38.0	24.2	38.3	24.1	38.2	24.1	38.6	24.0	37.8	23.9	38.9	25.6	38.6
12	52.2	25.9	52.3	25.8	52.2	25.8	51.6	25.9	51.0	25.6	52.1	25.9	54.0
14	63.3	24.4	63.4	24.3	63.3	24.4	62.9	24.6	62.4	24.5	63.7	24.1	64.1
16	72.2	23.2	72.2	23.2	72.2	23.2	71.6	23.4	71.1	23.1	73.1	22.8	74.0
total with Mom ed	55.9	27.7	56.0	27.6	55.9	27.6	55.6	27.8	54.7	27.4	58.8	27.3	

Edit 1: Includes language barrier cases, outliers, and low response PSU
Edit 2: Includes language barrier cases and low response PSU. Outliers are deleted
Edit 3: Includes language barrier cases. Outliers and low response PSUs are deleted
Edit 4: Includes language barrier cases and low response PSUs. Note that Age eligibility is age at date tested
Edit 5: Includes language barrier cases and low response PSUs. Highest grade completed is used in post-stratification weight

Distribution list

Annotated Briefing D0003839.A2

DMDC MONTEREY BAY CA
Attn: Dr. John Welsh (15 copies)
Dr. Daniel Segall
Ms. Kathleen Moreno
OASD (FMP) (MPP) (AP) WASHINGTON DC
Attn: Dr. Jane Arabian (8 copies)
Dr. Steve Sellman
DCNO (M&P) WASHINGTON DC
Attn: Mr. Ed Bres (N13T1)
HQ USMEPCOM CHICAGO IL
Attn: LCDR Leslie Turley (MOP-TD)
Mr. Rick Branch (MOP-TA)
HQMC M&RA QUANTICO VA
Attn: CAPT John America, USMC (MPP-50)
HQ U.S. COAST GUARD WASHINGTON DC
Attn: Ms. Mary Norwood (G-WTT-2)
DMPM-MPA WASHINGTON DC
Attn: Mr. James Call (DCS PERSONNEL)
HQ U.S. AIR FORCE WASHINGTON DC
Attn: Dr. Paul DiTullio (DPFPT)
RAND SANTA MONICA CA
Attn: Dr. Lawrence Hanser
NPRST MILLINGTON TN
Attn: Ms. Janet Held (PERS-13)
U.S. ARMY RESEARCH INSTITUTE ALEXANDRIA VA
Attn: Dr. Len White (PERI-RS)
HQ USAFRS/RSOAM RANDOLF AFB TX
Attn: Mr. George Germadnik (STE1)